Ratnesh Kumar

Delhi-110016, India | +91-9478752146 | ratneshcsit2021gate@gmail.com | LinkedIn | GitHub

SUMMARY

Data Scientist 4 year hands-on experience building and deploying production-level data pipelines and ML models. Developed and integrated advanced applications, including LLM-based chatbots and NLP search optimizations,while ensuring scalable and secure cloud infrastructure. Expertise in Python, SQL,ML,MLOps ,DL,GenAI and AWS Cloud Services, enabling efficient end-to-end project delivery and data-driven decision making.Hands-on experienced with large-scale data (1PB+).

TECHNICAL SKILLS

- **Programming Languages**: Python, SQL, PySpark, Linux Shell Scripting
- **Big Data & Streaming**: Apache Hadoop, Apache Spark, Apache Hive, Apache Airflow, Confluent Kafka, Databricks, BigQuery
- **Data Analysis & Visualization**: NumPy, Pandas, Matplotlib, SciPy, Seaborn, Plotly, Excel, Tableau, Power BI,QlikSense
- **Probability ,Statistics & Experimentation**: ANOVA, Hypothesis Testing, A/B Testing
- **Databases**: MySQL, PostgreSQL, MongoDB, Cassandra
- **Machine Learning & AI Models**: Supervised Learning (Linear Regression,Logistic Regression,Decision Tree, Random Forest,SVM,Naive Bayes,k-NN,Gradient Boosting Machines,XGBoost,LightGBM),Unsupervised Learning (k-Means Clustering,Hierarchical Clustering,DBSCAN ,PCA,t-SNE,Autoencoder,GMM), Time Series /Forecasting (ARIMA,SARIMA, Facebook Prophet)
- **Deep Learning Techniques**: Transformers, CNN, RNN, LSTM, GANs,OCR
- **Natural Language Processing:** NLTK, SpaCy, Gensim, TextBlob, BERT
- **Frameworks & Libraries**: Scikit-learn, TensorFlow, PyTorch, Keras, MXNet, OpenCV, Docling
- **Generative AI & Agentic AI**: GenAI Concepts, LLMs, Prompt Engineering, RAG, Fine Tuning LLMs (SFT, LoRA, QLoRA), Vector DB (FAISS,ChromaDB,Pinecone), LangChain, LangFlow, LangGraph,Autogen, OpenAI, Hugging Face,CrewAI, PhiData, Ollama,VertexAI,Groq, LlamaIndex
- **AWS Services:** S3, EMR, Lambda, DynamoDB, EC2, CloudWatch, IAM, Bedrock, Kendra, ECS, SageMaker, Redshift, Glue, ATHENA, SNS, SQS, QuickSight, CloudWatch, API Gateway,CloudFormation, Cognito,Guardrails, OpenSearch
- **MLOps**: AWS Ecosystem, MLFlow, Docker & DockerHub, Jenkins, Git/GitHub, GitLab, CI/CD, Streamlit, Pytest, Jira
- **Model Deployment & Monitoring**: Pipeline Deployment, End-to-End ML Systems, Evidently AI
- **Integration & API:** Developed and integrated RESTful APIs using FastAPI and FlaskAPI.

EXPERIENCE

TATA CONSULTANCY SERVICES **Dec 2021 - Present**

DATA SCIENTIST

Designed and maintained **government e-marketplace(https://gem.gov.in/)** data systems, deploying production-grade AI solutions and leveraging cloud services for real-time monitoring.

- **Agentic AI-Powered Price Comparison | Tech Stack- Agentic AI Framework- CrewAI, Platform AWS Cloud, S3 Bucket, Langchain, Python Libraries, AWS SageMaker, Custom Scraping tool using BeautifulSoup :** We developed a price comparison solution using two AI agents: the Search Logic Agent, which leveraged Agentic AI to dynamically generate search queries by concatenating the Golden Parameters (product specifications), Brand name, Product model and Product Title of products for a given category from the GeM database for precise product matching, and the Web Scraping Agent, which employed advanced scraping techniques to extract product price data from multiple e-commerce platforms (Flipkart, Amazon, Google Shopping). The entire solution was built as an end-to-end ML pipeline with production deployment on AWS SageMaker, applying coding best practices for modular, maintainable code and automated testing. Business Value -Other e-commerce does not showcase the price from other e-marketplaces to buyers. This solution helps buyers to take the decision to buy the product with a valuable price.
- **Agentic AI-Based Certificate Validation System | Tech Stack- Agentic AI Framework- CrewAI, Platform AWS Cloud, S3 Bucket, Langchain, Python Libraries, AWS Bedrock, AWS SageMaker, PyMuPDF :** Developed an Agentic AI-powered Certificate Validation Framework using the CrewAI framework to validate certificates (DPIIT, Trademark, MRP, FSSAI, Meity, PSARA, and BIS) uploaded by sellers in image or PDF format. The solution incorporated five intelligent agents: the TextExtractor Agent for preprocessing text data, the SectionSearchAgent to identify mandatory sections (e.g., logo, expiry date), the FuzzyMatcher Agent for fuzzy matching against a sample template using the Fuzzy Wuzzy library, and the ValidationAgent to analyze matching scores after retrieving seller information from the database and applying compliance checks. The system ensured documents adhered to predefined standards, such as logo presence, expiry date validity, and completeness, by checking for mismatches exceeding a 30% threshold and confirming all necessary sections were present. The final agent, the Comprehensive Validation Report Agent, generated a summary of the validation results. Business Value - Reduces 80% of the manual effort of the surveillance team which leads to saving of 5 PD(person days).

- **GeN AI Chatbot | Tech Stack- Platform AWS Cloud, KnowledgeBase Kendra, AWS Bedrock, LLM model Mistral-8x7B-Instruct, Lambda, Vector DB - Opensearch, Dynamo DB, Cloudwatch, CloudFormation, Cognito, API Gateway, S3 Bucket, Langchain, GitLab, GitHub Actions, Python Libraries, Embedding Model- Embed English v3 (transformer based) :**Ask GeM AI Developed and integrated a RAG-based GeM AI chatbot by leveraging GenAI concepts and AWS ecosystem services.Fine-tuned the Mixtral-8x7B-Instruct LLM on AWS Bedrock, integrated Amazon Kendra for semantic retrieval, and implemented hierarchical chunking for efficient long-document processing. Orchestrated secure storage with S3, real-time monitoring via CloudWatch, and logging with AWS Lambda to ensure scalable, reliable operation.Business Value -Significant reduction of 80% of contact center team.
- **BID document content validation and comparison | Tech Stack- AWS Bedrock, LLM Model- Opensource " Mixtral 8*7b Instruct ", S3 Bucket, GitLab, Github Actions, Python Libraries, Easy OCR, Docling, Langchain, vector DB- Pinecone, Embedding Model - "Embed English v3" by Cohere :** As part of the vendor assessment process, the customer surveillance team wants to validate the content of BID documents uploaded by the buyer and seller. Currently, this validation of the document requires manual intervention .Hence developed a solution to query the PDF file using generative AI LLM model. The solution invokes end users to upload their PDF documents and write their query. The system reads the PDF file using PyMuPDF and Docling, splits and stores the embedding into vector DB. The prompt will be entered by the end user in the form of query. The semantic search has been performed on vector DB to fetch the response of the user query. Business Value - Reduces 80% of the manual effort of the surveillance team which leads to saving of 5 PD(person days).
- **Improvisation of portal search functionality through vector DB | Tech Stack- Platform AWS Cloud, AWS Bedrock, LLM model Claude 3.5 Sonnet, Lambda, S3 Bucket, Kendra, Langchain, GitLab, Github Actions, Python Libraries, NLP (Natural Language Processing)** Optimized the backend search query performance by leveraging Vector Databases (Opensearch) to efficiently handle large e-commerce datasets. Integrated Generative AI techniques, including NLP processes that consist of SymSpell for advanced spelling correction and Verbosity for intelligent query refinement, to enhance the user search experience. Leveraged the ARPA probabilistic language model to improve query understanding and accuracy. Implemented Vector DB to embed and index complex e-commerce tables, enabling faster and more relevant search results.Prepared a vast corpus mapping Hinglish to their corresponding English words and correcting the misspelt english words. Developed a robust Flask web service to handle and process user search queries, ensuring efficient, scalable, and intelligent search functionalities. Business Value - The project reduced infrastructure costs by 30%, improved search speed by 50%, increased conversion rates by 20%, and enhanced operational efficiency by reducing query processing time by 40%.
- **AI-based Image Quality and Categorization Intelligence System | Tech Stack- Platform AWS Cloud, GitLab, Jenkins, Python Libraries (difPy, OpenCV, NumPy, Pandas), TensorFlow, CNN, Swin Transformer, Adam Optimizer, YOLOv8, Image Annotation Tool (Label Img) :** Developed an end-to-end ML system using machine learning models for image deduplication, relevance verification, and mismatch detection. Integrated a Flask API on the public sector e-commerce platform, leveraging AWS services for model training, deployment, and inference. Business Value - The project reduced manual review costs by 40%, minimized image-related listing errors, and saved operational costs by automating image classification and deduplication, improving overall efficiency.
- **Demand Forecasting (Time Series Forecasting and Demand Procurement Solution) | Tech Stack- Platform AWS Cloud, AWS SageMaker, Python Libraries, TSForecasting, DART Framework, S3 Bucket :** Developed a comprehensive time series forecasting solution as an end-to-end ML system, leveraging DART dataframe and ARIMA, SARIMAX, Exponential Smoothing models alongside Prophet. Implemented robust data preprocessing, model validation, and automated retraining pipelines on AWS SageMaker, ensuring reliable, scalable forecasts for order and shipment planning. Business Value - Minimized stockouts and overstocking, and improved procurement efficiency, leading to a 15% reduction in operational costs and more efficient resource allocation.
- **ML-Based Similar & Better Product Recommendation System | Tech Stack- Platform AWS Cloud, AWS SageMaker, Python Libraries, pandas, numpy, scikit-learn, S3 Bucket:** Cosine Similarity Identified similar and superior products using machine learning models deployed on the AWS ecosystem. Built a recommendation engine leveraging cosine similarity on feature embeddings, orchestrated continuous training and monitoring pipelines, and applied coding best practices for reliable model performance. Business Value - The project reduced product return rates by 15%, minimized inventory management costs, and improved sales efficiency by providing more accurate product recommendations, leading to a 10% increase in revenue.
- **Fraud Detection System | Tech Stack- Platform AWS Cloud, AWS Sagemaker, Python Libraries, pandas, numpy, scikit-learn framework, DBScan algorithm, Flask, PyMySQL, Kafka, PySpark, Amazon Quicksight** Developed a robust Bid Fraud Detection System for a Government e-commerce platform to detect potential collusion, monitor bid health, and identify irregular market behaviors. The solution utilized a data pipeline powered by Kafka messaging queues to receive and process bid information in real-time. In the first phase, advanced analytics were implemented to identify suspicious bidding patterns, including sellers and buyers sharing the same IP/MAC addresses, targeting specific buyer profiles, geographical clustering of bidders, and manipulation of bids at the last moment. The second part of the system focused on assessing the health of bids by comparing current bids against historical data, leveraging key parameters such as price variation comparison, seller type analysis, and qualification of sellers from past bids. This analysis enhanced the detection of collusion and market manipulation. The third phase, a real-time dashboard was created for the surveillance team, providing intuitive visualizations and alerts to monitor ongoing buyer and seller activities, thereby enhancing

the platform's integrity and transparency. This solution significantly improved market fairness and compliance by ensuring proper surveillance of the bidding process. Business Value - The project reduced manual monitoring costs by 4 PD, and increased process efficiency through automated real-time surveillance and fraud detection.

EDUCATION

WOOLF UNIVERSITY **Aug 2022 - 2024**

MS, Artificial Intelligence and Machine Learning -Online

•**GPA:** 4.0 / 4.0

GURU NANAK DEV ENGINEERING COLLEGE **Jul 2016 - 2020**

B.Tech, Information Technology

•**CGPA:** 7.34 / 10.0

PROJECTS

WIKIPEDIA PAGE VIEW FORECASTING FOR AD OPTIMIZATION

•Forecasted fluctuations in page visits for 145k Wikipedia pages over 550 days using SARIMAX and Facebook Prophet models.

•Achieved an overall MAPE of approximately 5.6% with exogenous variables.

Image Category Relevance | Tech Stack-YOLO, Python Libraries, Image Annotating Tool

• Object Detection with YOLO - Developed an Object Detection model using YOLOv8 to detect and categorize devices(e.g., computers, cameras, printers) in images. The model is trained on a custom dataset, system enabled real-time image prediction.The results were efficiently stored in a structured CSV format for streamlined reporting and data analysis.

ACHIEVEMENTS & CERTIFICATION

•GATE CS & IT Qualified (2020, 2021)

•Complete TensorFlow2 and Keras Deep Learning Bootcamp